International Research Journal on Advanced Engineering and Management
https://goldncloudpublications.com
https://doi.org/10.47392/IRJAEM.2024.0298
e ISSN: 2584-2854
Volume: 02
Issue: 06 June 2024
Page No: 2024-2028

Prediction of Liver Cirrhosis Using Classification Algorithms

Swedha[1], P. Rajesh[2], S. Muruganandham[3]
[1]PG, Department of Data Science, SASTRA Deemed University, Vadapalani, Chennai, India.
[2,3]Assistant Professor, Department of Data Science, SASTRA Deemed University, Vadapalani, Chennai, India.
Emails: *swedhashree26@gmail.com[1], itsrajesh91@gmail.com [2], muruga.anand.phd@gmail.com[3]*

Abstract
The liver is known as the largest internal organ of the body and is well known for its unique property of regeneration. One of the most common diseases of the liver is liver cirrhosis. Liver cirrhosis is one of the most widespread chronic diseases. It is characterized by the gradual replacement of the liver tissue with the scarring tissue. Liver cirrhosis, being asymptomatic, makes it difficult to identify the disease and diagnose it. Due to this factor, there is no cure for liver cirrhosis, but rather to prevent the spread of cirrhosis and reverse its effects. The main objective of this case study is to employ machine learning techniques to predict the probability of being affected by this disease. Various machine learning algorithms are employed for classifying whether a patient will develop liver cirrhosis or not [5-7]. By using various classification algorithms, the highest accuracy was achieved through Logistic Regression, XG Boost, and KNN. The accuracy was 81% for all three cases. Logistic regression, with an execution time of 0.094 seconds, outperformed the other two models. This shows that logistic regression provides a significantly more accurate prediction of liver cirrhosis disease.

Keywords: *Disease; Liver cirrhosis; Machine learning; Prediction*

1. Introduction
The liver, the second largest organ in the body and the largest internal organ, is responsible for the synthesis of albumin, prothrombin, fibrinogen, and antithrombin III, among many other essential bodily processes. It also secretes bile for digestion. Prothrombin and fibrinogen are necessary for blood clotting, whereas albumin stops fluid from leaking from blood arteries[10]. Antithrombin III controls coagulation to avoid the production of too many clots. Prolonged blood clotting times are caused by the liver's decreased ability to generate these proteins in cirrhosis. Chronic liver damage causes a condition called liver cirrhosis, in which scar tissue replaces healthy liver tissue. Chronic hepatitis, long-term alcohol abuse, and non-alcoholic fatty liver disease are all potential causes of this scarring. Hepatitis is an inflammation of the liver that can be brought on by autoimmune diseases, persistent infections with hepatitis B and C, and other conditions. Hepatitis is a blood-borne illness that spreads easily through bodily fluids, including blood [9][8]. Because liver cirrhosis is frequently asymptomatic in its early stages, early detection is difficult. On the other hand, the use of machine learning algorithms in medicine can improve the ability to diagnose long-term conditions like liver cirrhosis.

1.1. Statement of the Problem
Because liver cirrhosis is asymptomatic in the early stages, people find it challenging to recover from their condition. There is presently no known cure for liver cirrhosis. Its effects can be decelerated, but only if made carefully and gently. As such, there can be no guarantee that a procedure or medication that is advertised as curing liver cirrhosis. People experience lengthy waiting times even prior to diagnosis because of inadequate resources and diagnostics. Liver cirrhosis is currently regarded as the leading cause of death globally and poses the biggest risk to the human population. Diagnostic mistake also contributes significantly to the rising death rate since there are insufficient instruments available.

1.2. Aim and Significance of the Study
The study's objective is to put into practice a prediction model that can determine whether a patient will get liver cirrhosis in the early stages of the

International Research Journal on Advanced Engineering and Management
https://goldncloudpublications.com
https://doi.org/10.47392/IRJAEM.2024.0298
e ISSN: 2584-2854
Volume: 02
Issue: 06 June 2024
Page No: 2024-2028

condition. Three machine learning algorithms, including random forest and logistic regression, are trained on the patient data to determine their accuracy. Support vector machines (SVM) and forests. Given the patient's data, it ought to be able to determine whether the disease is present. Since liver cirrhosis is currently such a fatal condition, it must be lessened, which can only be done by receiving an early diagnosis. For the condition to be prevented and treated, an early diagnosis is crucial.

1.3. Literature Review

Jamila, G., Wajiga, G. M., Malgwi, Y. M., & Maidabara, A. H. (2022) report that Naïve Bayes, Classification and Regression Tree (CART), and support vector machine (SVM) were used in a suggested model with 10-fold cross-validation to predict hepatic cirrhosis. The model's performance was also evaluated through accuracy, precision, recall and F1 Score. Support Vector Machine outperformed the other two algorithms with an accuracy of 73%. [1] S.M. Abd El-Salam evaluated the performance of machine learning approaches on prediction of esophageal varices for Egyptian liver cirrhosis patients. The researcher has employed 14 algorithms like Support Vector Machine, Random Forest, Multilayer Perceptron, Naïve Bayes, and Bayesian Net. The highest performance was achieved by Bayesian Net with an accuracy of 68.9% [2]. Rahman S, Javed FM, Shamrat M, Tasnim Z, Roy J, Hossain SA (2019), conducted a comparison study on liver disease prediction using supervised machine learning algorithms. The researcher employed LR, DT, KNN, SVM, Naïve Bayes and RF. Logistic Regression gave the most accurate prediction with 75% accuracy. [3] Dritsas E, Trigka M (2023), used machine learning algorithms to predict the prevalence of liver disease. The researcher employed supervised machine learning models and ensemble techniques in the Indian Liver Patient's Records Dataset. With 10-fold cross-validation, the voting classifier outperformed other models with accuracy of 80%.[4]

1.4. Gaps in existing knowledge

Liver cirrhosis progresses through many phases. There are no instruments available to determine the patient's stage. Medical practitioners need to investigate how cirrhosis is progressing in the intended sample [11-13]. Lifestyle and genetic factors can be analysed to ascertain the onset and course of the illness. Once the patient's disease stage and severity have been determined, tailored treatment plans must be administered. For this reason, medical professionals must possess extensive knowledge and resources on the illness. This is only possible with a great deal of investigation. The necessary instruments for the disease's diagnosis must be created. Collaboration between computer scientists, medical specialists, and diverse health care providers can make this possible.

2. Methodology

The dataset came from the Mayo Clinic study on primary biliary cirrhosis (PBC) of the liver, which ran from 1974 to 1984.[14]. In all, 424 PBC patients were sent to the Mayo Clinic and other facilities over that ten-year period. The dataset's first 312 instances, which were included in the randomised experiment, have largely complete data. Even though they opted out of the clinical trial, the additional 112 cases consented to have their baseline measurements recorded and to be followed up on for survival. The results shown here comprise an additional 106 instances in addition to 312 randomly assigned individuals, since six of the cases were lost to follow-up shortly after diagnosis.

2.1 Data Pre-Processing

Data preprocessing is crucial before data analysis, involving the conversion of integrated data from various sources into a desired format [15]. Missing values are identified through statistical summaries, and their counts are displayed. Given the dataset's small size (418 rows), missing values are imputed rather than removed. For numerical data, the median is used due to its robustness against outliers. For categorical data, the mode is used. Categorical values are then converted to numerical values for further analysis. The dataset is split into training and testing data, with the Stage column as the target variable. Models are trained on the training data and tested on the testing data to evaluate accuracy [16].

2.2 Proposed Workflow

The data has been cleaned, prepared and employed in the model. The trained model is compared with other

International Research Journal on Advanced Engineering and Management
https://goldncloudpublications.com
https://doi.org/10.47392/IRJAEM.2024.0298
e ISSN: 2584-2854
Volume: 02
Issue: 06 June 2024
Page No: 2024-2028

models and the best performing model is chosen by comparison of the model. Figure 1 shows the workflow of the study.

2.3 Classification Algorithms used

The table 1 shows the various machine learning algorithms used to train the dataset. Machine learning algorithms, especially classification algorithms are used to train the dataset as the target variable, in this case, Stage, is a binary variable that consists of 0 or 1. 0 mentioning the absence of disease and 1 mentioning about the presence of the disease.

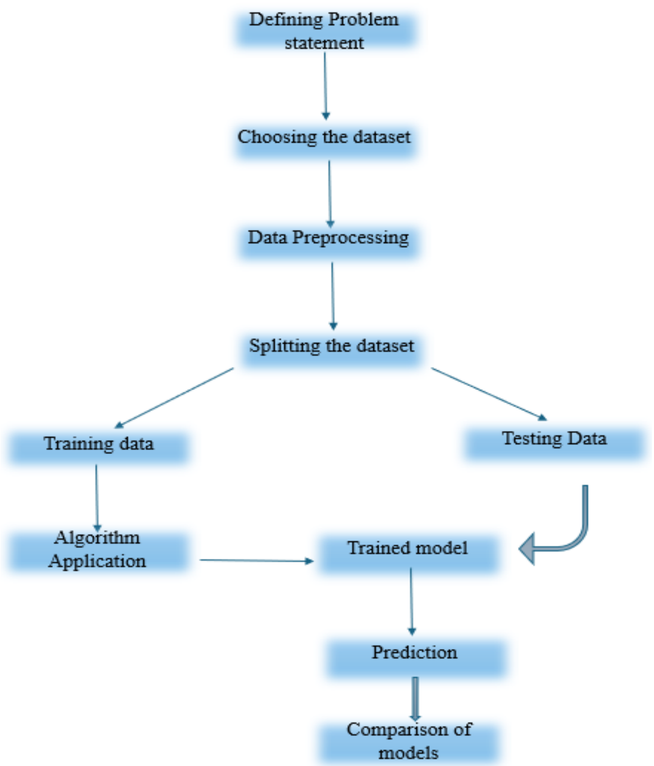

Figure 1 Workflow of the Study

Table 1 Algorithms Used

Algorithms Used	Accuracy
Logistic Regression	81%
Naïve Bayes	79%
K Nearest Neighbors	81%
XG Boost	81%
Support Vector Machine	79%

3. Results and Discussion

3.1.Results

The models that gave the highest accuracy in predicting the occurrence of the disease are Logistic Regression, XG Boost and KNearest Neighbors. All the models achieved the highest accuracy of 81%. Since all the models share the same accuracy, the execution time is noted as to seeing which model has given highest accuracy with less execution time. Comparing the execution time of all the models, Logistic Regression has a minimum execution time of 0.083 seconds. Table 2 shows the execution time of the three models. Hence Logistic Regression is the best performing model in accordance with accuracy and execution time.

Table 2 Execution Time of the Models

Models	Execution Time
Logistic Regression	0.094 seconds
XG Boost	0.244 seconds
K Nearest Neighbors	0.154 seconds

3.2.Evaluation of the models

Evaluation of the models is done using the confusion matrix and classification report. The classification report gives information about the accuracy, precision, recall, f1-score, and support of the model. Figure 2, figure 3, figure 4 shows the evaluation of the three best performing models.

```
              precision    recall  f1-score   support

           0       0.82      0.93      0.87        29
           1       0.78      0.54      0.64        13

    accuracy                           0.81        42
   macro avg       0.80      0.73      0.75        42
weighted avg       0.81      0.81      0.80        42

confusion matrix:
 [[27  2]
 [ 6  7]]

By using Logistic Model, the accuracy is 81%.
```

Figure 2 Classification report of LR

e ISSN: 2584-2854
Volume: 02
Issue: 06 June 2024
Page No: 2024-2028

```
Execution time: 0.24438786506652832 seconds
Accuracy: 0.8095238095238095
[[25  4]
 [ 4  9]]
Classification Report:
              precision    recall  f1-score   support

           0       0.86      0.86      0.86        29
           1       0.69      0.69      0.69        13

    accuracy                           0.81        42
   macro avg       0.78      0.78      0.78        42
weighted avg       0.81      0.81      0.81        42
```

Figure 3 Classification report of XG B

```
Execution time (in sec):  0.15427112579345703

K-Nearest Neighbors Accuracy: 0.8095238095238095
K-Nearest Neighbors Confusion Matrix:
[[27  2]
 [ 6  7]]
K-Nearest Neighbors Classification Report:
              precision    recall  f1-score   support

           0       0.82      0.93      0.87        29
           1       0.78      0.54      0.64        13

    accuracy                           0.81        42
   macro avg       0.80      0.73      0.75        42
weighted avg       0.81      0.81      0.80        42
```

Figure 4 Classification report of KNN

3.3. Future works

In future, when this model is trained with large amount of data from the clinical trials of liver cirrhosis patients, it will yield much more accuracy in predicting whether the patient will develop liver cirrhosis or not. Much research should be conducted in this domain, and this will help in the earlier diagnosis of the disease

Conclusion

For the given liver cirrhosis dataset, the best model that gave the maximum accuracy is Logistic Regression. Logistic regression gave a prediction with accuracy of 81% with least amount of execution time. Logistic regression is thus efficient for this dataset. Logistic regression has outperformed all the other algorithms in predicting the occurrence of Liver cirrhosis in a patient.

Acknowledgements

We acknowledge the Mayo Clinic for the liver cirrhosis dataset.

References

[1] Jamila, G., Wajiga, G. M., Malgwi, Y. M., & Maidabara, A. H. (2022). A diagnostic model for the prediction of liver cirrhosis using machine learning techniques. Computer Science & IT Research Journal, 3(1), 36–51. https://doi.org/10.51594/csitrj.v3i1.296

[2] Mukherjee, M., et al. "Improvement of the properties of PC/LCP blends in the presence of carbon nanotubes." Composites Part A: Applied Science and Manufacturing 40.8 (2009): 1291-1298.

[3] Ayyar, Manikandan, et al. "Preparation, characterization and blood compatibility assessment of a novel electrospun nanocomposite comprising polyurethane and ayurvedic-indhulekha oil for tissue engineering applications." Biomedical Engineering/Biomedizinische Technik 63.3 (2018): 245-253.

[4] Rajasekar, R., et al. "Development of compatibilized SBR and EPR nanocomposites containing dual filler system." Materials & Design 35 (2012): 878-885.

[5] Velu Kaliyannan, Gobinath, et al. "Influence of ultrathin gahnite anti-reflection coating on the power conversion efficiency of polycrystalline silicon solar cell." Journal of Materials Science: Materials in Electronics 31 (2020): 2308-2319.

[6] Rajasekar, R., et al. "Investigation of Drilling Process Parameters of Palmyra Based Composite." (2021).

[7] S. M. Abd El-Salam, M. M. Ezz, S. Hashem, W. Elakel, R. Salama et al., “Performance of machine learning approaches on prediction of esophageal varices for Egyptian chronic Liver cirrhosis patients,” Informatics in Medicine Unlocked, vol. 17, no. September, pp. 100267, 2019

[8] Rahman S, Javed FM, Shamrat M, Tasnim Z, Roy J, Hossain SA (2019) A comparative study on liver disease prediction using supervised machine learning algorithms. Int J

Sci Technol Res 8(11) [online]. Available www.ijstr.org

[9] Dritsas E, Trigka M (2023) Supervised machine learning models for liver disease risk prediction. Computers 12(1):19. https://doi.org/10.3390/computers12010019

[10] Hanif, I., & Khan, M. M. (2022). Liver cirrhosis prediction using machine learning approaches. 2022 IEEE 13th Annual Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON). https://doi.org/10.1109/uemcon54665.2022.9965718

[11] https://www.researchgate.net/publication/345846831_Classification_of_Hepatic_Disease_Using_Machine_Learning_Algorithms

[12] https://www.coursera.org/articles/machine-learning-in-health-care

[13] https://columbiasurgery.org/liver/liver-and-its-functions

[14] http://ncbi.nlm.nih.gov/12468243/#:~:text=The%20main%20nonparenchymal%20cells%20of%20the%20liver%2C%20Kupffer,their%20own%20proliferation%2C%20and%20effects%20on%20hepatocyte%20proliferation

[15] https://columbiasurgery.org/liver/liver-and-its-functions

[16] https://www.kaggle.com/datasets/fedesoriano/cirrhosis-prediction-dataset